Multivariate analysis

Introduction:

Scientific inquiry is an iterative learning process. Objective pertaining to the explanation of a social or physical phenomenon must be specified and then tested by gathering and analyzing data. In turn, analysis of the data gathered by experimentation or observation will usually suggest a modified explanation of the phenomenon. Throughout this iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigation to collect observations on many different variables.

This script is concerned with statistical methods designed to elicit information from these kinds of data sets. Because the data include simultaneous measurements on many variables, this body of methodology is called multivariate analysis.

Objectives:

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following:

1. Data reduction or simplification:

The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier.

2. Sorting and grouping:

Groups of similar objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.

3. Investigation of dependence among variables:

The nature of the relationship among variables is of interest. Are all variables mutually independent or are one or more variables dependent on the others? If so, how?

4. Prediction:

Relationship between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observation on the other variables.

5. Hypothesis construction and testing:

Specific statistical hypothesis, formulated in terms of the parameters of multivariate population, are tested. This may be done to validate assumptions or to reinforce prior convictions.

Applications of multivariate techniques:

The published applications of multivariate methods have increased tremendously in recent years. It is now difficult to cover the variety of real-world applications of these methods with brief discussions, as we did in earlier editions of this book. However, in order to give some indication of the usefulness of multivariate techniques, we offer the following short descriptions of the results of studies from several disciplines. These descriptions are organized according to the categories of objectives given in the previous section. Of course, many of our examples are multifaceted and could be placed in more than one category.

Data reduction or simplification:

• Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed.

(See Exercise 1.15)

• Track records from many nations were used to develop an index of performance for both male and female athletes.

(See [8] and [22].)

• Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions.

(See [23].)

• Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants.

(See [13].)

• A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediators judge the tactics they use in resolving disputes was determined. (See [21].)

Sorting and grouping:

• Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization.

(See [2].)

• Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from non-alcoholics.

(See [26].)

• Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease.

(See Exercise 1.14)

• The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not.

(See [31].)

Investigation of the dependence among variables:

• Data on several variables were used to identify factors that were responsible for client success in hiring external consultants. (See [12].)

• Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not.

(See [3].)

• Measurements of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations between pulp fiber properties and the resulting paper properties. The goal is to determine those fibers that lead to higher quality paper.

(See [17].)

• The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance.

(See [18])

Prediction

• The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college.

(See [10).)

• Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments.

(See [7] and [20].)

• Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.

(See [28].)

• cDNA microarray experiments (gene expression data) are increasingly used to study the molecular variations among cancer tumors. A reliable classification of tumors is essential for successful diagnosis and treatment of cancer. (See [9].)

Hypotheses testing:

• Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends.

(See Exercise 1.6)

• Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores.

(See [27].)

• Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories.

(See [16] and [25].)

• Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation.

(See [15])

The preceding descriptions offer glimpses into the use of multivariate methods in widely diverse fields.

The Organization of Data

Throughout this text, we are going to be concerned with analyzing measurements made on several variables or characteristics. These measurements (commonly called data) must frequently be arranged and displayed in various ways. For example, graphs and tabular arrangements are important aids in data analysis. Summary numbers, which quantitatively portray certain features of the data, are also necessary to any description.

We now introduce the preliminary concepts underlying these first steps of data organization.

Arrays:

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number $p \geq 1$ of variables or characters to record. The values of these variables are all recorded for each distinct item, individual, or experimental unit. We will use the notation x_{jk} to indicate the particular value of the kth variable that is observed on the jth item, or trial. That is,

x_{jk} = measurement of the kth variable on the jth item

Consequently, n measurements on p variables can be displayed as follows:

	Variable 1	Variable 2	⋯	Variable k	⋯	Variable p
Item 1:	x_{11}	x_{12}	⋯	x_{1k}	⋯	x_{1p}
Item 2:	x_{21}	x_{22}	⋯	x_{2k}	⋯	x_{2p}
⋮	⋮	⋮		⋮		⋮
Item j:	x_{j1}	x_{j2}	⋯	x_{jk}	⋯	x_{jp}
⋮	⋮	⋮		⋮		⋮
Item n:	x_{n1}	x_{n2}	⋯	x_{nk}	⋯	x_{np}

Or we can display these data as a rectangular array, called X, of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

The array X, then, contains the data consisting of all of .the observations on all of the variables.

Example 1.1 (A data array)

A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales) : 42 52 48 58

Variable 2 (number of books) : 4 5 4 3

Using the notation just introduced, we have

$x_{11} = 42$ $x_{2l} = 52$ $x_{3l} = 48$ $x_{4l} = 58$

$x_{12} = 4$ $x_{22} = 5$ $x_{32} = 4$ $x_{42} = 3$

and the data array **X** is

$$X = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

with four rows and two columns.

Considering data in the form of arrays facilitates the exposition of the subject matter and allows numerical calculations to be performed in an orderly and efficient manner. The efficiency is twofold, as gains are attained in both (1) describing numerical calculations as operations on arrays and (2) the implementation of the calculations on computers, which now use many languages and statistical packages to perform array operations. We consider the manipulation of arrays of numbers in Chapter 2. At this point, we are concerned only with their value as devices for displaying data.

Descriptive Statistics

Arrays of Basic Descriptive Statistics

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{\boldsymbol{x}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{x}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{x}}_{\boldsymbol{p}} \end{bmatrix}$$

Sample variances and covariances are

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & & s_{2p} \\ \vdots & & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample correlations

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & & r_{2p} \\ \vdots & & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Exercise:

Compute the basic descriptive statistics:

Eat	Appetite	Skin Reaction
2.222	1.945	1.000
2.312	2.312	2.000
2.455	2.909	3.000
2.000	1.000	1.000
2.727	4.091	2.000
2.000	2.600	2.000
1.875	1.563	0.000

2.388	4.000	2.000
2.273	3.272	2.000
2.000	1.000	2.000

Liquidity = amount of money

Compiled by: Abdullah Adil Mahmud